Cristina Fernández-Alcaina/Eva Fučíková/Jan Hajič/Zdeňka Urešová

The SynSemClass lexicon

A resource for multilingual synonymy

Keywords: Synonymy, Valency, Lexicon, Multilingual

This paper presents ongoing work on the multilingual lexicon SynSemClass (henceforth, SSC). Other related projects have addressed the issue of synonymy in multilingual contexts, e.g., EuroWordNet (Vossen 2004) or Predicate Matrix (Lopez de Lacalle et al. 2016). However, to the best of our knowledge, only SSC formalizes multilingual verbal synonymy in terms of syntactic and semantic properties. SSC is also linked to other resources in an effort to contribute to linked data in line with initiatives such as the *Unified Verb Index* (UVI)[1].

For our purposes, synonymy is understood as contextual synonymy (Palmer 1981) and context is defined in terms of the set of semantic roles expressed by the valency frame of a verb, either explicitly or implicitly and with possible restrictions. Based on these criteria, verbs are grouped into synonym classes, both monolingually and cross-lingually. Verbs are considered to belong to the same class if they convey the same meaning in a specific context, i.e., if the valency frame defined for each verb can be mapped to the set of roles (i.e., Roleset) of a class. For example, the Roleset defined for the class *allow* ('An Authority allows an Affected entity to engage in a Permitted entity') has three roles (Table 1). Each role is mapped to one of the arguments of the appropriate verb in each language, based on the information provided by the resources used. The same verb can be classified into different classes depending on the meaning expressed by their arguments (Fig. 1).

	Authority	**Permitted**	**Affected**
allow	ACT	EFF	PAT
dovolit	ACT	PAT	ADDR
erlauben	VA0	VA1	VA2
permitir	arg0	arg1	arg2

Table 1: Role-argument mapping in class *allow* (English, Czech, German and Spanish) (simplified)

'A Participant_1 meets a Partipant_2'		
	Participant_1	Participant_2
encounter	ACT	PAT
meet	**ACT**	**PAT**
sejít se	ACT	PAT
setkat se	ACT	PAT

'A Cognizer gets to know a Person'		
	Cognizer	Person
aquaint	ACT	PAT
meet	**ACT**	**PAT**
poznat	ACT	PAT
seznámit se	ACT	PAT

Fig. 1: An example of a verb (*meet*) included in two classes based on the different semantic roles expressed by the arguments it takes

[1] https://uvi.colorado.edu/ (last access: 4 May 2023)

SSC is built following a bottom-up approach and data are linked to a set of external resources available for each language (e.g., VerbNet for English or E-VALBU for German, among others). Translational equivalents are automatically extracted from parallel corpora (e.g., ParaCrawl for German-English) and annotated by human annotators. All annotators are (near-) native speakers of one of the languages included and proficient in English. For each language, the same set of classes is processed by two annotators and their annotations are monitored by a researcher. The task of the annotators consists in:

i) mapping the valency frame of a particular verb with the set of roles defined for the class where the verb is included as a potential class member,
ii) when available, establishing links to external resources, and
iii) selecting relevant examples.

The latest release, SynSemClass 4.0 (June 2022)[2] covering Czech, English, and German, contains 978 classes with approx. 9,000 class members (CMs). The Spanish-English part of the lexicon (planned to be included in the fifth version, 2023) contains 99 classes enriched by 620 Spanish class members (as of March 2023). Ongoing work is being done to include other languages in SSC.

The resulting resource has a twofold value: it provides fine-grained syntactic-semantic information on multilingual verbal synonyms at the same time it links data to other existing monolingual and multilingual resources. Although the number of classes and languages available in SSC is still limited, we believe that the resource can provide relevant data for descriptive and computational purposes as it may be used for cross-linguistic research on verbal valency as well as curated data for NLP tasks, such as cross-lingual synonym discovery.

References

Lopez de Lacalle, Maddalen/Laparra, Egoitz/Aldabe, Itziar/Rigau, German (2016): Predicate Matrix: automatically extending the semantic interoperability between predicate resources. In: Language Resources and Evaluation, 50(2), pp. 263–289. http://doi.org/10.1007/s10579-016-9348-5.

Palmer, Frank Robert (1981): Semantics. Cambridge University Press.

Vossen, Piek (2004): EuroWordNet: A Multilingual Database Of Autonomous And Language-Specific WordNets Connected via an Inter-Lingual-Index. In: International Journal of Lexicography, 17(2), pp. 161–173. http://doi.org/10.1093/ijl/17.2.161.

Contact information

Cristina Fernández-Alcaina

Institute of Formal and Applied Linguistics (Charles University in Prague

alcaina@ufal.mff.cuni.cz

Eva Fučíková

Institute of Formal and Applied Linguistics (Charles University in Prague)

[2] https://lindat.cz/services/SynSemClass40/ (last access: 4 May 2023)

fucikova@ufal.mff.cuni.cz

Jan Hajič

Institute of Formal and Applied Linguistics (Charles University in Prague)

hajic@ufal.mff.cuni.cz

Zdeňka Urešová

Institute of Formal and Applied Linguistics (Charles University in Prague)

uresova@ufal.mff.cuni.cz